1994

The Reliability and Validity of an Object Assembly Task

Dane'l A. Koonce
Eastern Illinois University

THESIS REPRODUCTION CERTIFICATE

TO: Graduate Degree Candidates (who have written formal theses)

SUBJECT: Permission to Reproduce Theses

The University Library is receiving a number of requests from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow theses to be copied.

PLEASE SIGN ONE OF THE FOLLOWING STATEMENTS:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

[redacted]

Author

3-29-96

Date

I respectfully request Booth Library of Eastern Illinois University not allow my thesis to be reproduced because:

Author

Date

The Reliability and Validity

of an Object Assembly Task

(TITLE)

BY

Dane'l A. Koonce

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Specialist in School Psychology

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

1994

YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

7-19-94

DATE

[redacted]

7-19-94

DATE

[redacted]

DEPARTMENT HEAD

List of Tables

TABLE OF CONTENTS

Acknowledgements

I wish to thank the four members of the committee who assisted me with this thesis. Without the direction and efforts of Drs. Rick Cross, Julia Jorgenson, William McGown and Fred Yaffe, this study would not have been possible. I wish to extend a special thanks to Shannon Mast and David Workman who assisted me in the initial stages of data collection.

This thesis is dedicated in memory of Steven Wall, who introduced me to computers and challenged me to pursue my educational goals.

Abstract

A computerized object assembly task (COA) was constructed and compared to the Wechsler Adult Intelligence Scale-Revised (WAIS-R) object assembly subtest (OA). The two tasks, the COA and the OA, were administered to seventy-one college undergraduates. The correlation between the WAIS-R OA subtest and the two COA conditions, Preview and NoPreview, combined was moderate and statistically significant. The findings support the feasibility of constructing a computerized version of the OA task.

Chapter 1

Introduction

The purpose of this thesis is to construct a computerized object assembly task and compare it to the WAIS-R Object Assembly subtest.

The history of computerized testing dates back to the 1960s. One of the earliest attempts to automate psychological testing was the Automated Pictoral Paired and Associate Learning Task (APALT), which was developed by Gedye and his colleagues (Gedye, 1965, 1967a, b, 1970; Gedye and Miller, 1969; Geyde & Wedgewood, 1966). The APALT development was in response to a need for a simple and objective measure of clinical progress in elderly patients with dementia (Wedgewood, 1982). Another major contributor to the field of automated psychological testing was David Weiss and his co-workers. In collaboration with the Office of Naval Research, Kiely, Zara, and Weiss (1900) developed an interactive time-sharing system. This system was developed as a research tool, controlling the administration of a range of tests using a number of different process control strategies, including the use of branching rules to select the items that are administered to the subject. In a similar manner, these large, time-sharing computers were used to scan, score, and profile standardized tests, and to provide an interpretive report (Sampson, 1983). However, the impact of automated testing was limited to non-interactive procedures (Thompson & Wilson, 1982).

These limitations shifted with the introduction of the cathode-ray tube (CRT) and software packages that were adapted for the administration of psychological tests. The CRTs were then connected by telephone lines or by satellites, allowing remote users, i.e. test takers, to answer individual test items via a computer terminal keyboard. The automation of traditional paper-and-pencil tests, in conjunction with the technological advances in microcomputers, led to the adaptation of psychological tests for handicapped individuals as well as able-bodied individuals (Maguire, Knobel, Knobel, Sedlacek, & Piersel, 1991; Sampson, 1983).

Currently, there are a number of assessment systems which administer, score, and interpret tests on site and provide immediate feedback through computer analysis (Brown, 1984). Psyc-Systems, in Baltimore, Maryland, is the principal manufacturer of large integrated testing systems and has been developing testing software for time-shared computer systems since the mid-1970's. Most of the software packages that are available from Psyc-Systems have been adapted from paper-and-pencil instruments during the last 20 years. A few of the more popular software packages that Psyc-Systems has available are the Minnesota Multiphasic Personality Inventory, California Psychological Inventory, Visual Search Task and Eysenck Personality Questionnaire.

The presentation format for most of the instruments is either true/false or multiple choice, but a few of the instruments are more interactive. For example, the Medical History, Self-Directed Search, and Social History software packages use conditional logic routines, which vary the type of response based on the individual responses of the client (Brown, 1984). Other instruments, such as the Speilberger State Trait Anxiety Scale, Jenkins Activity Survey, and the Eysenck Personality Questionnaire, are simply a computer-administered version of the traditional format, for which item presentation is the same for each individual.

Based on the possibilities that new technology offers, it appears that more school psychologists are using computers as a part of their daily routines for report writing, data analysis, and assessment purposes. Jacob and Brantly (1989) surveyed 268 school psychologists to explore the nature and extent of legal-ethical problems associated with computer applications in school psychology. They found that 37% of those surveyed use computers for data storage, 35% for scoring verification, 33% for test interpretation, and 32% for report writing. The use of interactive testing was not included as part of the survey.

The rise in computer utilization has not been universally heralded as an advance in the practices of psychological testing. Altemose and Williamson (1981) have argued that the increasing level of computerized assessment could lead to the replacement of school psychologists by less competent psychometricians utilizing computers.

Regardless of the criticisms, computers are useful instruments for stimulus presentations and data acquisition, as well as in the administration, scoring, and interpretation of psychological tests. The microcomputer's flexibility allows for adaptive or tailored methods of administration. Furthermore, researchers are finding that computerized assessment offers more benefits, such as the attainment of more objective and standardized testing procedures, and the flexibility to assist individuals with auditory, visual, and physical limitations. Space (1981) recognized that computerized psychological testing "frees the psychologist from repetitive tasks to spend additional time on more complex considerations, such as psychodiagnosis beyond standard computer-retrieved information, psychotherapy, community preventive work, and research" (p. 598).

Benefits of computerized assessment

Computer-administered psychological tests offer several advantages over the traditional format. For example, computerized versions offer an improved standardization of test administration, improved speed and accuracy of administration, scoring compilation of results, and improved cost effectiveness (Bartram & Bayliss, 1984; French 1986; Hasselbring, 1984). Some investigators have found that when the computerized version was correlated with the conventional form of the test, no significant differences were found in rank order; they also were found to be significantly and positively correlated (Knight, Richardson, & McNarry, 1973; Overton & Scott, 1972; Wilson, Thompson, & Wylie, 1982). The current literature supports the use of computers in this context, and the research thus far indicates that no major differences between the standard administration and computerized version exist.

A computerized battery of psychological tests, used in conjunction with a software package that provides automated scoring, interpretation, and profile analysis, can reduce the turn-around time between the completion of testing and the return of the report (Space, 1981). In addition, Johnson and Williams (1990) found that the cost of a computerized evaluation was close to half the cost of a traditional test battery (equipment costs excluded).

Another benefit that automated psychological testing can offer is tailored or adaptive testing. In adaptive testing, the computer program adjusts the test difficulty in relation to the performance of the individual being tested. Consequently, the administration of unnecessary items is reduced, and the duration of the testing session is ultimately shortened. Hulin, Drasgow, and Parsons (1983) pointed out that in addition to a reduction in testing time, fatigue, and boredom, adaptive testing can increase measurement accuracy. Moreover, Weiss (1985) asserted that since adaptive testing is response-contingent, it has been able to yield measurements of comparable or superior quality to those of conventional tests with considerably fewer items administered to each individual. Additionally, Weiss (1985) stated that the increases in testing efficiency can be attributed to the reduction of administration time, making it possible to measure two or more traits in the same amount of time that would be required to measure a single trait using conventional tests, which translates into higher degrees of reliability and potentially higher levels of validity.

One of the most significant advantages of computerized assessment is its ability to adapt to individuals with auditory, visual and physical limitations in the testing population. The microcomputer can be paired with specialized input and output devices that provide individuals with handicaps the opportunity to complete various tests with minimal assistance (Sampson, 1983). Thus, its unique capabilities allow for opportunities in research that may have been formerly difficult or impossible.

<u>Problems associated with computerized assessment</u>

Although these benefits are practical and far-reaching, some authors (Hofer & Green, 1985; Greaud & Green, 1984; Burke & Normand, 1987) have acknowledged that some potential problems exist in this area of testing. One of the main criticisms that faces computerized psychological testing is that even though high correlations exist between the paper-and-pencil tests and their computerized versions, they do not provide sufficient evidence for demonstrating equivalence (Burke & Normand, 1987). A finding that may partially explain the imperfect correlation is the way in which the task must be modified in order for it to be automated. Thus, the type of response required by the examinee (i.e., auditory, via keyboard, touch sensitive screen, or mouse) on the computerized version of the test may prevent the generalization of certain psychometric properties (i.e., norms, reliability, and validity) of the conventional version (Burke & Normand, 1987).

Because computer-oriented tests are available to a wide range of personnel, another source of concern is the development of standards for computer-based psychological testing. Many of the professionals--including psychologists, social workers, family and child therapists, physicians, nurses, and business and personnel managers--likely to use the tests may lack necessary education and training (Matarazzo, 1986; Groth-Manet & Schumaker, 1989; Sampson, 1983). Therefore, there is a necessity for professional organizations to adopt a set of standards for the utilization of computerized testing instruments. In 1986, the American Psychological Association (APA) began reevaluating their standards for computerized assessment, based on the earlier work of Hofer and Bersoff, in order to develop an acceptable set of professional standards. The guidelines set forth by the APA further emphasized that practitioners should have an adequate knowledge base of the instrument chosen, which included familiarity with psychological measurement, background in the history of the test being used, research of the tests, and knowledge of the area of intended applicability (Groth-Manet & Schumaker, 1989).

A final concern is the possibility that the norms obtained from the conventional test may not be appropriate for use with the computer-tested subpopulations that they are applied to. If the computerized version of a conventional paper-and-pencil test fails to produce either comparable reliability, validity, or cutting score data, then the normative data cannot be generalized from the conventional test to the computerized counterpart.

Comments and criticisms of computerized assessment

The increasing popularity of computers in the field of psychology has influenced the opinions of many authors and researchers who support or oppose their use in assessment. Specifically, the area of equivalence has drawn the most attention. Lord (1980) noted that "frequency distributions of test scores in which no change in examinee's rank is observed between the conventional and computerized versions provides more sound evidence for equivalence" (p. 154). Beaumont (1981) has a more rigid standard that "only if the automated version performs as if it is a parallel form of the original version of the test will the norms established for the original version be appropriate" (p. 431). Beaumont (1985) added, "there has been as yet little systematic study of how the presentation of test materials by computer, or how the use of various response media, affect the outcome of the assessment procedure" (p. 11). Further examination of the implications of using the computerized counterpart in place of the conventional form of the test is necessary.

Interface factors

The standard administration of a computer-based task may alter the nature of the task due to inherent computer-linked factors and thereby affect the test's reliability and/or validity. These factors are the stimulus presentation, response format, and examiner presence. This thesis will focus on the joint effect of stimulus presentation and response format.

Stimulus presentation is the computer's ability to present designs or pictures on a monitor at their highest resolution. A poor stimulus presentation can disguise the examinee's true level of ability since it may affect performance efficiency and motivation on certain items. French & Beaumont (1990) stated that, stimulus degradation (i.e., a computer graphic with poor resolution) can cause an examinee to attend more fully to the degraded stimuli and, therefore, take more time and be more accurate. Poor stimulus presentation can also cause the subject to be less motivated to tackle the item and feel less capable of completing the task (French, 1986).

Response format is the interface factor that relates to the type of response device used by the subject to interface with the computerized version of a psychological test. Whether the interface is via a keyboard, touch screen, or mouse, the use of an unfamiliar response device as well as the awkwardness that the device might impair the subjects' performance on certain types of tasks (Burke & Normand, 1987). Ultimately, the subject's lack of familiarity with the interface can play a significant role in altering the test performance as well as impede its level of equivalence with comparable tests. The effects of either stimulus presentation or response format, or their joint effect may produce an effect on the covariation between the computerized version of the test and its non-computerized counterpart.

Review of Literature

There are several well known psychological tests that have been automated for use with normal populations and individuals with visual, auditory and physical handicaps, including the Raven's Standard Progressive Matrices (Calvert & Waterfall, 1982), portions of the WAIS-R (Wechsler, 1981), Matching Familiar Figures Test (Kagan, Rosman, Day, Albert, & Phillips, 1964), PPVT (Dunn, 1965) and the PPVT-R (Dunn & Dunn, 1981). Elwood and Griffen (1972) compared the automated version of the Wechsler Adult Intelligence Scale (WAIS) with the standard administration of the test which revealed reliability and validity coefficients in the .90's. Similar findings were reported for the computer administrations of the Minnesota Multiphasic Personality Inventory (Lushene, O'Neal, & Dunn, 1972) and the Raven Progressive Matrices Test (Hitti, Riffer, & Stuckless, 1971). Overton and Scott (1972) compared the hand-administered to the automated version of the Peabody Picture Vocabulary Test with a large sample of individuals with mental retardation and reported very high correlations (in the range of .91 to .94).

A majority of the evidence on computer adaptations of paper-and-pencil questionnaires points to the tentative conclusion that non-equivalence is typically small enough to be of no practical consequence (Moreland, 1985). Rezmovic (1977) found that computer administration caused extreme scores to become even more extreme, implying that non-equivalence in questionnaires may occur at points in the distribution of scores where measurement is already imprecise. These findings appear to be due to the change in the response format between computer administered and conventional versions of the test (Moreland, 1985). In addition, if caution is exercised to insure that the two administration formats are as similar as possible, this problem disappears (Biskin & Kolotkin, 1977; Bresolin, 1984). Practitioners must ensure that before inferences and generalizations are made from the computer counter-part of an existing paper-and pencil test, the observed means, variances, and correlations between the two versions are nearly equal (Allen & Yen, 1979; Hofer & Green, 1985). One reasonably can ask whether the development of a computer adapted psychological test can accurately measure the verbal and/or performance component of a standardized intelligence test.

Van Merrienboer, Jelsma, Timmermans, & Sikken, (1989) compared a computerized version of the Matching Familiar Figures Test (MFFT) to the traditional, experimenter-controlled MFFT. In their small sample of undergraduate students, no differences were found between internal consistencies and test-retest reliabilities of the computerized version and those of the traditional form of the MFFT. The internal consistency coefficients for the standard and the computerized version of the MFFT were medium for errors, and high for response times.

The authors recognized that three of the possible factors that may account for the differences in mean test scores were the degradation of stimuli, the presentation of the pictures on the computer screen, and a novelty effect of computerized testing (Van Merrirnboer et al., 1989). In fact, the degraded stimuli may cause some subjects to attend to the stimuli more, take more time, and, thus, increase their accuracy. The novelty effect may also have been a

factor that accounts for the differences in mean test scores. The subjects were not familiar with the mouse-controlled interfaces, and support for this explanation is offered in the stability data. After the second administration of the computerized version, the novelty effect decreased and the differences in the mean test scores were far less pronounced.
The investigators concluded that the computerized version of the MFFT was at least as reliable as the standard form, taking the low reliability of the original test into account, and there was no strong reason to believe that the tests do not measure the same construct. Finally, the investigators observed a reflexive behavior (that is, gathering information more systematically and carefully when attempting to solve a problem and the solution is not immediately obvious) on the computerized version of the MFFT, which may be explained as a novelty effect and, possibly, as an effect of the degradation of stimuli.

French and Beaumont (1990) conducted a clinical study of the automated assessment of intelligence by the Mill Hill Vocabulary test and the Standard Progressive Matrices test compared to the standard versions of these tests. A total of 184 subjects were assessed using the Standard Progressive Matrices test while 129 were retested on the automated version. The mean scores and standard deviations for the group on the first test administration did not differ significantly between the keyboard and the touch-screen versions. The test-retest reliability was high.

After the data from the retested subjects were submitted for analysis of variance with keyboard vs. touch-screen, standard-first vs. computer-first, and standard score vs. computer score as main effects, the analysis revealed two significant main effects and no interactions. The first main effect showed that the computer score was significantly lower than the standard score. The second main effect noted that the subjects who received the standard version first scored significantly higher on the Standard Progressive Matrices test than those who were administered the computerized version of the test first. The authors explained that the difference in scores between the standard and computerized versions was probably due to poor resolution of the computer's graphic system. The significance of this factor is worth comparing to the results found by Calvert and Waterfall (1982) and Watts, Baddeley, and Williams (1982), both of whom used systems with higher graphic qualities. The results obtained by Calvert and Waterfall indicated a nonsignificant trend for subjects to score higher on the standard compared to the automated version, and in the Watts et al. study, subjects obtained a significantly higher score on the standard version of the Standard Progressive Matrices test. The results from these studies imply that some factors other than stimulus quality played a role in altering the performance on the computerized test in the present study (French & Beaumont, 1990).

The poor stimulus presentation was noticeable on several items and would severely hamper the performance of subjects with poor eyesight. They noted, "subjects who received the computerized version first and performed at less than their true level of ability on certain items, due to poor graphics, presumably felt less motivated to tackle the same items on the clearer standard version because they had convinced themselves that these items were too difficult for them" (p. 138). The issue of screen resolution was a concern throughout the study, yet its effect was not anticipated to be dramatic. Consequently, the absolute difference in scores between the standard and computerized versions does not permit the routine use of the computerized test in place of the standard.

Martin (1989) is the only person actively publishing articles about the usefulness of HyperCard for various types of psychological testing research. Specifically, his work concentrates on the HyperCard administration of a WAIS-R Block Design subtest. After the development of the block design stackware, Martin and Wilcox (1989) administered the computerized block design task in addition to the Wechsler Adult Intelligence Scale-Revised (WAIS-R) subtest with the standard materials provided in a WAIS-R kit to undergraduates that were enrolled in two introductory psychology classes at a small liberal arts university. The split-half reliability and validity coefficients were calculated for each condition of the administration to determine the amount of covariance that exists between the computer administered task and the WAIS-R Block Design subtest.

Martin and Wilcox (1989) reported that the elapsed time on the computer administered task was found to be a reliable measure, which correlated moderately with the elapsed time on the WAIS-R Block Design subtest. Additionally, accuracy was reliable, but its correlation with accuracy on the WAIS-R subtest failed to reach significance. They identified two possible explanations for the failure to achieve a significant validity coefficient for correctly completed designs. The low correlations may be attributed to the restriction of range and commission of different types of errors during the two tasks. An error analysis revealed that of the 66 errors committed during the stackware task, all but one was due to the subject failing to complete the design within the established time limit. On the computerized version of the WAIS-R designs, 25 errors were the result of running out of time while 14 errors were committed as a result of subjects reporting the completion of the design, and then noticing that they had misplaced a block.

After completion of the initial study (Martin, 1989), Martin and Allen (1992) replicated the study using a nontemporal approach to scoring the task and modifying the software to increase the difficulty of the computer task. The difficulty of the computer task was increased due to "the high percentage of designs that were completed by the subjects in both tasks, and the substantially lower number of errors committed on the computer version of the task" (p. 1). By eliminating the four-block designs and exclusively utilizing nine-block designs, the ceiling effect in the data was eliminated, which may have compromised the correlation with the WAIS-R subtest (Martin & Juniper, 1992).

At the completion of this study, reliability coefficients were high (r_{xx} = .72) for the WAIS-R Block Design subtest and (r_{xx} = .97) for the computer administered block design task. Furthermore, the correlation between the computer block design task and the WAIS-R subtest was (r_{xy} =.60). The authors also found that when looking at the combined effect of score with time, the subjects took twice as long to complete the computerized designs; however, a high percentage of the designs were correctly completed by subjects in both tasks. Thus, the study established that moderate reliability can be achieved with a nontemporal measure of proficiency on a computer block design task.

In their exploration of a nontemporal measure of performance, Martin and Allan did not construct a scaling device, which awards bonus points for quick perfect performances at the completion of each item. For three of the WAIS-R tests (Arithmetic, Block Design, and Object Assembly), the raw scores reflect both quality and speed of response (Wechsler, 1981). When the solution to an item is reached quickly, bonus points are added to the item score, which is intended to increase the variability of scores on a test and to improve its reliability. The addition of time bonuses to Martin and Wilcox's initial study may have increased the low correlation that resulted from restriction of range and commission of errors.

Purpose of the Study

This study proposes to construct and then compare a computer-based object assembly test (COA) with the Object Assembly (OA) subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R). The principal means of comparison will be to correlate the scores of the OA and COA task. Item analysis on the COA provides data from the item tryout to determine the degree to which performance on one item correlates with the collective performance on all the other items in the criterion test, and identifies both the items that measure the construct and those that can discarded or altered (Allen & Yen, 1979).

The items selected should maximize the criterion-related validity. The final step is to determine whether the COA task can be considered quasi-Tau Equivalent, where variances and correlations with criterion variables are similar between the two forms of the test, to the OA subtest of the WAIS-R. This is assessed by regressing the OA subtest across the COA, and by establishing the hit rate of the COA against the OA subtest.

Hypotheses

It is expected that the COA and OA will correlate at least moderately. However, the factors of response format and stimulus presentation may have the effect of impairing some subjects' performance due to a lack of familiarity with the computer itself or the response medium. On the WAIS-R subtest, the number of puzzle pieces does not determine the level of difficulty for the puzzle, which may alter the nature of the task for the computerized version. In a like manner, a time constraint is also likely to affect the subjects' performance on the task. Martin and Wilcox (1989) noted that elapsed mean times differed substantially between the computer-administered task and the WAIS-R subtest. They found that subjects completed the WAIS-R subtest more rapidly than the computerized version, and most of the errors committed on the automated version were due to failing to complete the design within the established time limit. The elapsed time and the number of correctly assembled puzzles for the computerized version is important for establishing the degree of convergence between the two conditions and serves as a measure of performance. However, based on the work of Martin (1992) and Van Merrienboer et al. (1989), a temporal score may be less susceptible to variance introduced by changes in the computer environment, such as porting the task to other machines, the use of degraded stimuli, and use of different input devices.

Finally, it is suspected that in a computerized testing situation, the performance of individuals who are unfamiliar with computers may be impaired. Studies have linked computer unfamiliarity and anxiety (Hedl, O'Neil, & Hansen, 1973; Russon, Josefowitz, Edmonds, 1994) and their combined effect has been associated with lower test performance on computerized tests (Johnson & White, 1980; Lee, 1986; Lee, Moreno, & Sympson, 1986) and significantly higher achievement scores (Chin, Donn, Conroy, 1991).

Chapter 2

Method

Subjects

Seventy-one college students (59 females and 12 males) who were enrolled in undergraduate Psychology courses at Eastern Illinois University, in central Illinois participated in the study. The subjects were given extra credit by their instructor for participating in the research. Subjects ranged in age from 18 to 44; the mean age was 23.5.

Materials

The Wechsler Adult Intelligence Scale-Revised (WAIS-R) object assembly subtest was administered with the standard materials provided in a WAIS-R kit. The apparatus used to administer the computerized block design task was a Apple Macintosh Centris 610 computer with a monitor that has a maximum pixel dimension of 640 x 870 pixels, and a standard Apple mouse. The stackware with which the computerized version of the object assembly task will be administered using the HyperCard 2.0 scripting language (Goodman, 1990) and a modified version of the HyperCard Puzzle program.

Procedure

The COA experimental puzzles were constructed from clip art and single line drawings (see Appendix A and B). The clip art and drawings were then divided into different sized configurations (i.e., 2 x 2, 2 x 3, 2 x 4, 3 x 3, and 4 x 4 puzzles) based on the size of the clip art or drawings. These COA puzzles were then placed in order ranging in difficulty from simple to more complex, based upon the number of pieces. After the first 20 subjects completed the COA experimental puzzles, a time bonus routine was developed to scale each item configuration based on a fast and perfect performance. Because three of the WAIS-R tests include bonuses for quick perfect performance, a time bonus routine was developed for the COA in order to increase the variability of scores on the test.

The scaling was completed by dividing the quartiles for each configuration (i.e., 2 x 2, 2 x 3, 2 x 4, 3 x 3, and 4 x 4) into an upper and lower range. These ranges were then assigned bonus points, based on the subject's time of completion for the COA item, with the maximum number being 8 points and lowest being 2 points. Bonus points were only assigned if all the pieces were placed correctly in the time allotted. If the subjects time of completion was past the allotted time, they only received points based on the number of correctly placed pieces at the cutoff time. Following the item analysis, the remaining 51 subjects were administered both the WAIS-R OA subtest and the revised version of the COA task. The time and the number of correct responses on each item for each subject were stored on data cards in the HyperCard stack for analysis of errors and time of completion. The other fifty-one subjects were randomly assigned to a counterbalanced manner for administration of the WAIS-R Object Assembly subtest and the COA. Group 1 consisted of 25 subjects who were given the WAIS-R Object Assembly subtest first and the COA task second. Group 2 consisted of 26 subjects who were given the COA task first and the WAIS-R Object Assembly subtest second. The testing lasted approximately 50 minutes. Prior to the administration of the COA and the OA, the subjects were interviewed to collect demographic information, discover their computer experience, and determine whether they had been given a WAIS-R in the past 6 months. The subject's had a mean of 5.18 years of computer experience (SD = 3.2, range = 0-15). The mean educational level was 15.74 years (SD = .83, range 14 - 16). After the interview, the subjects were given 10 sample items of the COA test in order to increase their proficiency with the use of a mouse.

During the COA task, the subjects were presented with 21 test puzzles. At the beginning of each trial, subjects either saw a three second display of the completed puzzle before the pieces were dispersed in a narrow area of the screen (Appendix A) or they saw only the pieces of the unassembled puzzle (Appendix B). When the subjects completed the puzzle, they were reminded to press the button marked "Finished" and the button marked "Start Puzzle" in order to proceed to the next puzzle. Each subsequent item increased in complexity and size (i.e., 2 x 2, 2 x 3, 2 x 4, 3 x 3, and 4 x 4) as the subjects proceeded through the experimental puzzles.

Analysis of Data

The standard scoring procedure for comparing scores for each puzzle on the WAIS-R subtest was a measure of performance. For the OA subtest, "the score for each item is equal to the number of cuts correctly joined, plus a maximum of 3 bonus points per item for quick, perfect performance" (Wechsler, 1981. p. 79). A quick and perfect performance refers to all the cuts for the pieces of the object that are correctly joined plus the bonus points assigned to the time of completion. The COA task was quantified by calculating the performance of each subject in three different conditions: COA Preview raw, COA NoPreview raw, and COA Combined raw. The assembly score for each of the COA conditions was calculated by counting the number of correctly placed pieces plus a maximum of 8 bonus points per item. After the sums were computed, the items within each COA condition generated the mean performance for each subject. The raw score for the four WAIS-R OA subtest items was calculating by using the standard scoring procedure stated previously. The scores for the four WAIS-R OA items were then used to generate each subjects mean performance.

Multiple regression analysis was conducted to support the criterion related validity and to determine whether the Preview or NoPreview items of the COA differentially correlated with the criterion measure. The hit rates, the proportion of the total population that are considered successful or unsuccessful in completing the tasks introduced (i.e., the COA conditions, and the WAIS-R OA subtest) based on a preestablished cutoff score, were also computed. Lastly, the partial correlations were computed to determine the contribution of the Preview condition toward the prediction of WAIS-R OA task partialling out the NoPreview condition and then determining the contribution of the NoPreview condition by partialling out the Preview condition.

Chapter 3

Results

Table 1 presents all the raw score means, standard deviations, and correlations for the all of the COA items with the total raw score of the WAIS-R OA subtest. Inspection of the Pearson-product moment correlations among the COA items and the total raw score of the WAIS-R OA subtest showed that of the twenty-one COA items only four were moderately related to the WAIS-R OA subtest. Correlations among the COA items and the WAIS-R OA subtest range from .47 to .08.

Criterion-Related Validity

Pearson product-moment correlations between the best overall Preview, NoPreview and the WAIS-R OA subtest total scores (presented in Table 2) were moderate to low. The correlations among the best overall Preview and NoPreview items ranged from .47 to .29. Table 3 presents the correlations between the best preview COA items with the total raw score of the WAIS-R OA subtest. The correlations ranged from .47 to .21. Table 4 presents correlations for the best overall NoPreview COA items with the total raw score of the WAIS-R subtest.

Table 1

Means, Standard Deviations, and Correlations of the COA items with the WAIS-R Object Assembly items

Item[a]	M[b]	SD	Correlation with WAIS-R OA subtest
1_8	12.1	3.4	.42
2_8	9.5	4.0	.35
3_16	15.4	5.5	.43
4_8	8.1	3.7	.15
5_6	12.2	2.6	.16
6_9	8.1	2.9	.21
7_8	12.6	2.8	.08
8_9	12.6	3.0	.13
9_8	13.4	2.7	.32
10_9	10.7	3.7	.28
11_6	10.2	3.2	.25
12_8	13.1	2.9	.14
13_8	11.5	3.5	.47
14_8	9.5	3.7	.24
15_8	9.6	3.6	.14
16_9	13.3	2.7	.29

Table 1 continued

17_9	12.9	3.3	.21
18_8	11.6	3.0	.40
19_16	14.2	5.6	.33
20_8	10.7	2.6	.20
21_9	10.2	3.7	.21

Note. The first number of the item refers to the order of administration. The second number refers to the puzzles configuration. 1_8 = 2 X 4; 3_16 = 4 x 4; 6_9 = 3 X 3. The means and standard deviations refer to the joint number of components assembled by the subjects based on the time allotted for each puzzle configuration.

Table 2

Descriptive Statistics for the Best Overall Combined COA Items

Item	M	SD	Correlation with the WAIS-R OA Subtest
1_8	12.1	3.4	.42
2_8	9.5	4.0	.35
3_16	15.4	5.5	.43
9_8	13.4	2.7	.32
13_8	11.5	3.5	.47
16_9	13.3	2.7	.29
18_8	11.6	3.0	.40
19_16	14.2	5.6	.33

Table 3

Descriptive Statistics for the Best Preview COA Items

Item	M	SD	Correlation with the WAIS-R OA Subtest
1_8	12.1	3.4	.42
3_16	15.4	5.5	.43
9_8	13.4	2.7	.32
11_6	10.2	3.2	.25
13_8	11.5	3.5	.47
17_9	12.9	3.3	.21
19_16	14.2	5.6	.33
21_9	10.2	3.7	.21

Table 4

Descriptive Statistics for the Best No Preview COA Items

Item	M	SD	Correlation with the WAIS-R OA Subtest
2_8	9.5	4.0	.15
4_8	8.1	3.7	.21
6_9	8.1	2.9	.28
10_9	10.7	3.7	.24
14_8	9.5	3.7	.29
16_9	13.3	2.7	.40
18_8	11.6	3.0	.20
20_8	10.7	2.6	.35

Table 5

Descriptive Statistics and Correlation Coefficients for the COA Items Overall

Variable	M	SD	Correlation with the WAIS-R OA subtest	r^2	Hit Rate
BPCOA	99.7	22.2	.52	.28	.65
BPCOA[a]	-	-	.47	.22	-
BNPCOA	81.3	18.4	.48	.23	.78
BNPCOA[b]	-	-	.37	.14	-
BP&NPCOA	121.6	27.2	.53	.28	.66
WAIS-R OA	31.4	4.4	-	-	-

Note.

BPCOA = Best preview COA items without outlier, BPCOA[a] = Best preview COA items with outlier, BNPCOA = Best no preview COA items without outlier, BNPCOA[b] = Best no preview COA items with outlier, BP&NPCOA = Best preview and no preview COA items, WAIS-R OA = Wechsler Adult Intelligence Scale-Revised Object Assembly subtest

Table 6

Full Model Simultaneous Regression of WAIS-R OA on BPCOA and BNPCOA

Variable	slope	partial-correlation	p-value
BPCOA	.08	0.11	.08
BNPCOA	.02	-0.01	.08

The correlations between the best NoPreview items and the WAIS-R OA subtest total score ranged from .20 to .40. Table 5 presents the correlations between the WAIS-R OA subtest and the COA items based on their group characteristics (i.e., best Preview items, best NoPreview items). The Pearson product-moment correlations revealed that the COA items grouped according to their characteristics, the BPCOA, BNPCOA and the BP&NPCOA, were at least moderately correlated to the WAIS-R OA subtest. Although their individual correlations were not significantly different, the Best-Preview (BPCOA) condition displayed a stronger relation with the WAIS-R OA task then the NoPreview (BNPCOA) condition. The examination of the correlations of the BPCOA[a] and BNPCOA[b] conditions (Table 5) show that an outlier can make a significant change in the magnitude a correlation. The r^2 change between the BPCOA and the BPCOA[a] was a difference that accounted for 6% of the variance. The same problem is also evident in the BNPCOA and the BNPCOA[b]. The r^2 change between the BNPCOA and the BNPCOA[b] was a difference of 9% of the variance. The difference between the Pearson product-moment of the BNPCOA and the BNPCOA[b] was also significant. The hit rate for the BP&NPCOA was .66, meaning that 66% of the test population was successful on both the Best-Preview and NoPreview items and the WAIS-R OA task based median cutoff scores for each test. The hit rate for the BPCOA was .65, meaning that 65% of the test population was successful on both the Best-Preview items and the WAIS-R OA task. Finally, the hit rate for the BNPCOA was .78, meaning that 78% of the test population was successful on the NoPreview items and the WAIS-R OA task.

The regression of the WAIS-R OA subtest on the BPCOA and the BNPCOA combined showed a moderate multiple correlation of .53, which was statistically significant $F_{.05}(2,47) = 9.45$ ($p < .0004$). Parameter estimates are provided in Table 6.

Chapter 4

Discussion

The results on the comparison between the summary performance on the WAIS-R OA, and the summary performance, on the COA, in all three forms, Best-Preview, Best NoPreview, and Best combined, would support the proposition that a moderate relationship exists between these tasks. The Preview condition displayed a somewhat more robust relationship with the WAIS-R OA task, in comparison to the NoPreview condition, but the relative superiority was not large. We then have the preliminary support for the feasibility of constructing a computerized version of the OA task which, under the proper scaling may lend itself to Tau-equivalence with the original WAIS-R.

During the development of the COA, there was some concern that the Preview condition of the COA task might act as a cued memory in the original stimulus presentation. As a result, the two different conditions for previewing were established, with no explicit theoretical rationale for including one procedure over the other, other than to note that the WAIS-R OA task is a "non-preview" test. Perhaps, the efficiency with which the subjects' produce an object out of parts that may not be immediately recognizable can be attributed to the underlying intellectual process, in addition to visual organization involved in the WAIS-R OA task.

In the standardized form of the WAIS-R task, test protocol prevents the test subject from examining the array while they are being arranged. These procedures essentially control for the time component in the recognition of the puzzle under assembly. In the Preview condition of the COA, there is by presumption, support that recognition of the object itself is not a factor in the intellectual process, since the subject sees the total assembly before proceeding to put its pieces together. The multiple regression sheds light on the difference between the Preview condition and the NoPreview condition; both conditions in combination adequately correlate with the WAIS-R OA subtest, but in a partially correlated condition, the NoPreview condition seemed rather inferior to the Preview condition (though the Preview condition did not quite make significance when partialled with the NoPreview condition.) This would at least suggest that the shared variance between the WAIS-R OA task and the COA is not due to a "recognition" factor. However, that a "recognition" factor is not central to the variance of the WAIS-R task is completely counter-intuitive. So it is quite likely that the variance of the COA is not a "central" variance factor with respect to intelligence.

At the onset of the study, it was hypothesized that the subjects performance would be impaired by a lack of computer familiarity or anxiety towards the use of a computer. It was concluded that the subjects performance may have been impaired by their lack of proficiency with the use of a mouse. After the administration of the sample COA items, most of the subjects had no difficulty in becoming accustomed to the use of the mouse. Although most subjects were not experienced with the use of a mouse, it is reasonable to say that other methods of interface between the user and computer might have allowed the subjects to perform at their optimal level.

Given the results of this study, it is recommended that the use of COA task should only be used as a research device. Limitations of this study support this recommendation, especially in terms of the programming environment of HyperCard. One limitation of this study is that the programming capabilities of the HyperCard stackware limited the variation of the size and shape of the puzzle pieces. All the puzzle pieces were constrained to the shape of squares, making it difficult to increase the complexity of the puzzle without adding additional pieces to the puzzles. Another limitation of this study is the small, restricted sample. Most of the subjects were females from rural communities, and of caucasian decent. With a sample makeup as this, it is impossible to generalize the findings to a larger more heterogeneous population.

Despite the limitations involved in this study, the possibilities for designing an object assembly task using other programming languages are more far-reaching. With further development of computer software, the possibility for the computerized assessment will likely be more feasible.

References

Allen, M. J. & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole

Altemose, J. R., & Williamson, K. B. (1981). Clinical judgement vs. the the computer: Can the school psychologist be replaced by a machine? Psychology in the Schools, 18, 356-363.

Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present, and future. Journal of Occupational Psychology, 57, 221-237.

Beaumont, J. G. (1981). Microcomputer-aided assessment using standard psychometric procedures. Behavior Research Methods and Instrumentation, 13, 430-433

Beaumont, J. G. (1985). The effect of microcomputer presentation and response medium on digit span performance. International Journal of Man-Machine Studies, 22, 11-18.

Biskin, B. H., & Kolatkin, R. L. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. Applied Psychological Measurement, 1, 543-549.

Bresolin, M. J., Jr. (1984). A comparative study of computer administration of the Minnesota

Multiphasic Personality Inventory in an inpatient psychiatric setting. Unpublished doctoral dissertation, Loyola University of Chicago.

Brown, D. T. (1984). Automated assessment systems in school and clinical psychology: Present status and future directions. School Psychology Review, 13, 455-460.

Burke, M. J., & Normand J. (1987). Computerized psychological testing: Overview and critique. Professional Psychology: Research and Practice, 18, 42-51.

Calvert, E. J., & Waterfall, R. C. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. International Journal of Man-Machine Studies, 17, 305-310.

Chin, C. H., Donn, J. S., & Conry, R. F. (1991). Effects of computer based tests on achievement, anxiety, and attitude of grade 10 science students. Educational and Psychological Measurement, 51, 735-745.

Cohen, R. J. (1985). Computer-enhanced qualitative research. Journal of Advertising Research, 25, 48-52.

Dunn, L. (1965). The Peabody Picture Vocabulary Test-Revised. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1981). The Peabody Picture Vocabulary Test-Revised. Circle Pines, MN: American Guidance Service.

Elithorn, A., Morington, S., & Stavrou, A. (1982). Automated psychological testing: Some principles and practice. International Journal of Man-Machine Studies, 17, 247-263.

Elwood, D. L., & Griffin, R. H. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. Journal of Consulting and Clinical Psychology, 38, 9-14.

French, C. C. (1986). Microcomputers and psychometric assessment. British Journal of Guidance and Counseling, 14, 33-45.

French, C. C., & Beaumont, J. G. (1990). A clinical study of the automated assessment of intelligence by Mill Hill Vocabulary test and the Standard Progressive Matrices test. Journal of Clinical Psychology, 46, 129-139.

Gedye, J. L. (1965). A Pictorial Paired and Associate Learning Test (PPA-1) [Film strip], 1st ed. London: Educational

System.

Gedye, J. L. (1967a). A Pictorial Paired and Associate Learning Test (PPA-2) [Film strip], 2nd ed. London: Educational System.

Gedye, J. L. (1967b). A teaching machine programme for use as a test of learning ability. In D. Unwin & J. Leedham (Eds.), Aspects of Educational Technology. London: Methuen.

Gedye, J. L. (1970). The use of an interactive computer terminal to access the effect of Cyclandelate on the mental ability of geriatric patients. In G. Stoker, R. A. Kuhn, P. Hall, G. Becker, E. Van Der Veen (Eds.), Assessment of Cerebrovascular Insufficiency: An international symposium on various methods of assessing the effect of Cyclandelate on higher cortical function, Wuzburg 1970 (pp. 29-42). Stuttgart: Georg Thieme Verlag.

Gedye, J. L., & Miller, E. (1969). The automation of psychological assessment. International Journal of Man-Machine Studies, 1, 237-262.

Gedye, J. L., & Wedgewood, J. (1966). Experience in the use of a teaching machine for the assessment of senile mental changes. Proceedings of the 7th International

Congress of Gerentology (pp. 205-207). Vienna.

Goodman, D. (1990). The complete HyperCard 2.0 handbook (3rd ed.). New York: Bantam Books.

Greaud, V. A., & Green, B. F. (1984). Equivalence of conventional and computer presentation of speed tests. Manuscript submitted for publication.

Groth-Marnet, G., & Schumaker, J. (1989). Computer-based psychological testing: Issues and guidelines. American Journal of Orthopsychiatry, 59, 257-263.

Hasselbring, T. (1984). Computer-based assessment of special needs students. Special Services in the Schools, 1, 7-19.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. (1973). Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40, 217-222.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. Journal of Consulting and Clinical Psychology, 53, 826-838.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Jacob, S., & Brantly, J. C. (1987). Ethical-legal problems with computer use and suggestions for best-practices: A national survey. School Psychology Review, 16, 69-77.

Johnson, D. F., & White, C. B. (1980). Effects of training on computerized test performance in the elderly. Journal of Applied Psychology, 65, 357-358.

Johnson, E. S., & Meade, A. C. (1987). Developmental patterns of spatial ability: An early sex difference. Child Development, 58, 725-740.

Johnson, J. H., & Williams, T. A. (1990). Using on-line computer technology to improve service response and decision-making effectiveness in a mental health admitting system. In J. B. Sidcowski, J. H. Johnson, & T. A. Williams (Eds.), Technology in mental health care delivery systems. Norwood, NJ: Albet

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflexive attitudes. Psychological Monographs, 78 (1, Whole No. 578).

Kiely, G. L., Zara, A. R., & Weiss, D. J. (1983). Alternative forms reliability and concurrent validity of adaptive and conventional tests with military recruits. San Diego, CA: Navy Personnel Research and

Development Center.

Knight, R. M., Richardson, D. H., & McNarry, L. R. (1973). Automated vs. clinical administration of the Peabody Picture Vocabulary test and Coloured Progressive Matrices. _American Journal of Mental Deficiency_, _78_, 223-225.

Lee, J. A. (1986). The effects of past computer experience on the computerized aptitude test performance. _Educational and Psychological Measurement_, _46_, 727-733.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. _Educational and Psychological Measurement_, _46_, 467-474.

Lord, F. M. (1980). _Applications of item response theory to practical testing problems_. Hillsdale, NJ: Erlbaum.

Lushene, R. E., O'Neil, H. H., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. _Journal of Personality Assessment_, _38_, 353-361.

Maccoby, E., & Jacklin, C. (1974). _The psychology of sex differences_. Stanford, CA: Stanford University Press.

Maguire, K. B., Knobel, M. M., Knobel, B. L., Sedlacek, L. G., & Piersel, W. C. (1991). Computer-adapted

PPVT-R: A comparison between standard and modified versions within an elementary school population. Psychology in the Schools, 28, 199-205.

Martin, T. A. (1990, August). Non temporal measures of performance on a computerized block design task. Paper presented at the 98th Annual Meeting of the American Psychological Association, Boston.

Martin, T. A., & Allen, W. E. (1992, August). Efficiency as an automated measure of block design performance. Paper presented at the Continental Convention of the American Psychological Association, Washington, DC.

Martin, T. A., & Juniper, M. D. (1993, August). Relationship of a computer block design task to the WAIS-R. Paper presented at the 101st Convention of the American Psychological Association, Toronto.

Martin, T. A., & Wilcox, K. L. (1989). HyperCard administration of a block design task. Behavior Research Methods, Instruments, and Computers, 21(2), 312-315.

Matarazzo, J. D. (1986). Response to Fowler and Butcher on Matarazzo. American Psychologist, 41, 96.

Morland, K. L. (1985). Computer-assisted psychological assessment in 1986: A practical guide. Computers in

Human Behavior, 1, 221-233.

Overton, G. W., & Scott, K. G. (1972). Automated and manual intelligence testing: Data on parallel forms of the Peabody Picture Vocabulary test. American Journal of Mental Deficiency, 76, 639-643.

Rezmovic, V. (1977). The effects of computerized experimentation on response variance. Behavior Research Methods and Instrumentation, 9, 144-147.

Russon, A. E., Josefowitz, N., Edmonds, C. V. (1994). Making computer instruction accessible: Familiar analogies for female novices. Computers in Human Behaviors, 10, 175-187.

Sampson, J. P. (1983). Computer-assisted testing and assessment: Current status and implications for the future. Measurement and Evaluation in Guidance, 5, 293-299.

Space, L. G. (1981). The computer as psychometrician. Behavior Research Methods and Instrumentation, 13, 595-606.

Van Merrienboer, J. J. G., Jelsma, O., Timmermans, J., & Sikken, J. (1989). Computerized vs. experimenter controlled administration of the Matching Familiar Figures Test. Educational and Psychological

Measurement, 49, 883-891.

Thompson, J. A., & Wilson, S. L. (1982). Automated psychological testing. International Journal of Man-Machine Studies, 17, 279-289.

Van Merriener, J. G., Jelsma, O., Timmerman, J. & Sikken, J. (1989). Computerized vs. experimenter controlled administration of the Matching Familiar Figures Test: Mean test scores and reliabilities. Educational & Psychological Measurement, 49, 883-892.

Watts, K., Baddeley, A., & Williams, M. (1982). Automated tailored testing using Raven's Matrices and the Mill Hill Vocabulary tests: A comparison with manual administration. International Journal of Man-Machine Studies, 17, 331-344.

Wechsler, D. (1981). Wechsler Adult Individual Scale-Revised. San Antonio, TX: Psychological Corporation.

Wedgewood, J. (1982). The Automated Pictoral Paired and Associate Learning Task. International Journal of Man-Machine Studies, 17, 291-296

Weiss, D. J. (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53, 774-789.

Wilson, S. L., Thompson, J. A., & Wylie, G. (1982). Automated psychological testing for the severely handicapped. International Journal of Man-Machine Studies, 17, 291-296.